University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2020

Experimenting with Generative Adversarial Networks to Expand Sparse Physiological Time-Series Data

Baumgartner, Martin ; Eggerth, Alphons ; Ziegl, Andreas ; Hayn, Dieter ; Schreier, Günter

Abstract: Machine Learning research and its application have gained enormous relevance in recent years. Their usage in medical settings could support patients, increase patient safety and assist health professionals in various tasks. However, medical data is often sparse, which renders big data analytics methods like deep learning ineffective. Data synthesis helps to augment small data sets and potentially improves patient data integrity. The presented work illustrates how Generative Adversarial Networks can be applied specifically to small data sets for enlarging sparse data. Following a state-of-the-art analysis is conducted, experimental methods with such networks are documented, which have been applied to three different data sets. Results from all three sets are presented and take-away messages are summarized. Concluding, the results' quality and limitations of the work are discussed.

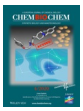
DOI: https://doi.org/10.3233/SHTI200103

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-195910
Journal Article
Published Version

The following work is licensed under a Creative Commons: Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

Originally published at:
Baumgartner, Martin; Eggerth, Alphons; Ziegl, Andreas; Hayn, Dieter; Schreier, Günter (2020). Experimenting with Generative Adversarial Networks to Expand Sparse Physiological Time-Series Data. Studies in Health Technology and Informatics, 271:248-255.
DOI: https://doi.org/10.3233/SHTI200103

ChemPubSoc Europe
DOI: 10.1002/cbic.201900346
CHEMBIOCHEM Full Papers



Identification of Synthetic Activators of Cancer Cell Migration by Hybrid Deep Learning

Dominique Bruns[+],[a] Erik Gawehn[+],[a] Karthiga Santhana Kumar,[b] Petra Schneider,[a] Martin Baumgartner,[b] and Gisbert Schneider*[a]

Deep convolutional neural networks (CNNs) are a method of choice for image recognition. Herein a hybrid CNN approach is presented for molecular pattern recognition in drug discovery. Using self-organizing map images of molecular pharmacophores as input, CNN models were trained to identify chemokine receptor CXCR4 modulators with high accuracy. This machine learning classifier identified first-in-class synthetic CXCR4 full agonists. The receptor-activating effects were confirmed by intracellular cAMP response and in a phenotypic spheroid invasion assay of medulloblastoma cell invasion. Additional macromolecular targets of the small molecules were predicted in silico and tested in vitro, revealing modulatory effects on dopamine receptors and CCR1. These results positively advocate the applicability of molecular image recognition by CNNs to ligand-based virtual compound screening, and demonstrate the complementarity of machine intelligence and human expert knowledge.

Introduction

Machine learning models have become a cornerstone of computer-assisted drug discovery. Among the many different approaches, neural networks constitute a particularly active field of research. Deep convolutional neural networks (CNNs) were initially developed for image recognition and have recently been adopted by the life sciences,[1,2] for example, for rapid and robust microscopic and angiographic image analysis,[3] and pattern recognition in genome data.[4] CNNs excel at feature extraction from images, which has been evidenced not only by great success in computer vision, but also by promising use in medical imaging and radiotherapy.[1,2] In this study, CNN technology was used for ligand-based drug discovery. The method combines CNN-based pattern recognition with self-organizing maps (SOMs)[5] for representing molecular structures as standardized images. The results of a prospective application provide proof-of-concept for this concept of collaborative machine intelligence.[6] The new hybrid deep learning method enabled the identification of the first-in-class, synthetic low molecular weight activators of chemokine receptor CXCR4.

The CXCR4 protein is a member of the G protein-coupled receptor (GPCR) family.[7] Intracellular signaling is triggered by the endogenous agonist CXCL12, a chemokine protein with a molecular mass of 8 kDa, and involves both G-protein activation and the recruitment of β-arrestin with subsequent receptor internalization.[8] CXCR4 can undergo dimerization to form homodimers or heterodimers with ACKR3, leading to differential and complex signaling regulation.[9,10] The CXCR4 receptor is constitutively expressed,[11] plays a key role in HIV infections[12,13] and has been identified as an anticancer target for drug discovery. Several CXCR4 antagonists have been published[14–16] including the marketed drug plerixafor. However, finding novel CXCR4 modulators, especially synthetic low-molecular-weight agonists as tool compounds that mimic the CXCL12 chemokine, has been proven difficult. Here, we used hybrid deep learning for virtual screening of a large screening compound collection to find innovative chemokine receptor modulators.

Results and Discussion

The first step of the deep learning approach was to develop a technique for converting chemical structures to two-dimensional images for further processing by the CNN (Figure 1). For this purpose, we extracted 495 827 bioactive compounds with annotated nanomolar activities from the ChEMBL23 database[17] and represented them in terms of their topological pharmacophore features (chemically advanced template search, CATS).[18] This process encoded each molecule as a 210-dimensional "CATS descriptor" representation. Molecules with similar CATS descriptors (i.e., similar topological shape and pharmacophore features) were then clustered on a two-dimensional grid using Kohonen's unsupervised SOM algorithm.[19]

[a] D. Bruns,[+] E. Gawehn,[+] Dr. P. Schneider, Prof. Dr. G. Schneider
Swiss Federal Institute of Technology (ETH)
Department of Chemistry and Applied Biosciences, RETHINK
Vladimir-Prelog-Weg 4, 8093 Zürich (Switzerland)
E-mail: gisbert@ethz.ch

[b] Dr. K. S. Kumar, Dr. M. Baumgartner
Paediatric Neuro-Oncology Research Group
Department of Oncology, Children's Research Center
University Children's Hospital Zürich
Lengghalde 5, 8008 Zürich (Switzerland)

[+] These authors contributed equally to this work.

Supporting information and the ORCID identification numbers for the authors of this article can be found under https://doi.org/10.1002/cbic.201900346.

© 2019 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

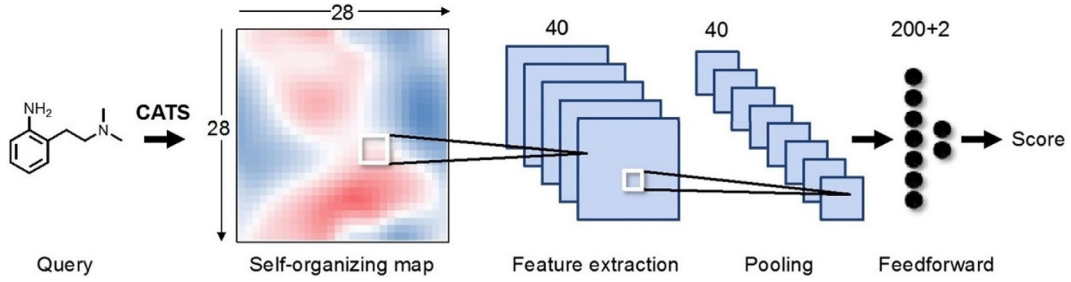


Figure 1. Hybrid deep learning architecture. Compounds were represented in terms of CATS pharmacophore descriptors, and converted into SOM images with a resolution of 28×28 pixels. Coloring (temperature scale) of the map pixels shows the local activation of the prototype pharmacophore patterns learned by the SOM. These molecular images were used as input to the CNN classifier. The best-performing CNN contained a convolution layer with 40 feature maps, a max-pooling layer, and a fully connected feedforward classifier with 200 + 2 neurons. The output value (*Score*) can be interpreted as the *pseudo*-probability of belonging to the positive training set (here: CXCR4 ligands).

The SOM identifies prototype patterns in the training data and arranges the resulting data clusters as a neighborhood-preserving map, where each grid point corresponds to one of the prototype patterns forming the cluster centroids.[20] The resulting map was used to generate images of the molecular representations (CATS).[21] Different strategies to leverage the information contained in SOM patterns have been proposed.[22,23] Here, we used a CNN for feature extraction from the SOM images (Figure 1) and applied the model to ligand-based virtual compound screening. The trained CNN model predicts ligand bioactivity based on the SOM excitation patterns (images) evoked by the input molecular structures.

The motivation to perform deep learning on top of molecular descriptor encoding and similarity clustering was threefold: Firstly, there is no best generic representation of a molecule for drug design.[24] Secondly, combining human and machine intelligence enables deep learning in low-data situations, taking advantage of the expert knowledge ingrained in the CATS descriptor, which is known to enable scaffold hopping in combination with machine learning.[18] Thirdly, artificial neural networks are not as flexible as the human mind when it comes to abstraction from very few input examples.[25] This means that learning essential known molecular pharmacophore features or variants thereof cannot be ensured when the underlying ligand data is scarce, as is often the case when one attempts to find compounds against new targets.[26] The method presented here aims to leverage prior background information in a semi-supervised fashion and enable deep learning in such situations.

The machine learning model was developed in two steps. First, a SOM was trained on a set small bioactive molecules, as described.[21,27] Then, different CNN classifiers were trained and tested for their ability to predict CXCR4 ligands, using the trained SOM model as input layer. The best CNN architecture contained an input layer with 28×28 = 784 neurons (i.e., the number of SOM clusters), a convolutional layer with a receptive field size of 4×4 pixels, *stride* = 1 and 40 feature maps, followed by a pooling layer using max-pooling with a window size of 2×2 pixels and *stride* = 1, and finally a feedforward layer with 200 input neurons and two output neurons, one for each data class (Figure 1). For prospective application, this deep network architecture was trained on all available data (392 CXCR4 modulators) and used for CXCR4 target prediction of the screening compound pool. (Figure 2, Table 1)

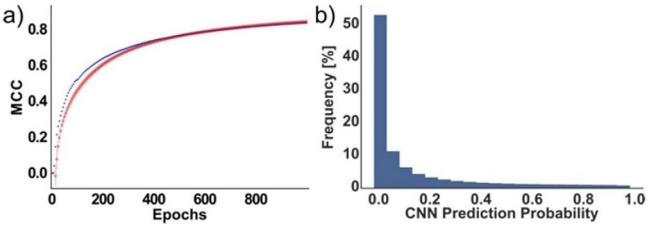


Figure 2. Convolutional neural network (CNN) training and prediction score distribution. A) Development of mean and standard deviation of the Matthews correlation coefficient (MCC, in [−1,1] with MCC = 1 indicating perfect prediction) for the best network architecture during cross-validation (red), and the MCC for training on the complete CXCR4 data (blue). B) Distribution of the predicted *pseudo*-probability of the CNN (*Score*) obtained for a library of 5.7 million screening compounds. The top-ranking compounds (*Score* > 0.995) were considered for bioactivity testing.

Table 1. Statistical evaluation of the deep learning classifier model. Mean and standard deviation of the Matthews correlation coefficient (MCC), accuracy, precision, recall and the receiver-operator characteristic area under the curve (ROCAUC, in [0,1] with ROCAUC = 1 indicating perfect prediction) both in five-fold cross-validation and after training the model on all CXCR4 data (final model). Results were collected as streaming metrics, meaning that they were updated continuously during the learning process.

	Cross-validation		Final model
	Training data set	Validation data set	Complete data set
MCC	0.84 ± 0.01	0.88 ± 0.04	0.87
accuracy	0.92 ± 0.01	0.94 ± 0.02	0.93
precision	0.92 ± 0.01	0.93 ± 0.03	0.93
recall	0.92 ± 0.01	0.95 ± 0.02	0.94
ROCAUC	0.92 ± 0.01	0.94 ± 0.02	0.93

Of the ≈5.7 million compounds in the screening compound pool, 7423 were predicted as active against CXCR4 with a network *Score* > 0.995. These potentially active molecules were

© 2019 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

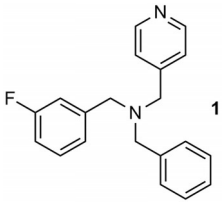
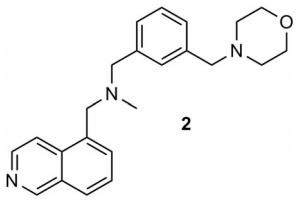
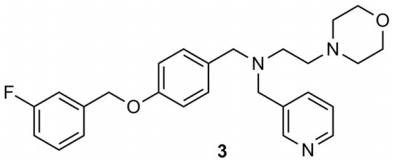
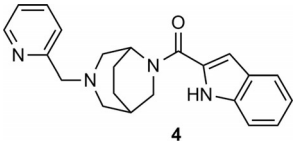
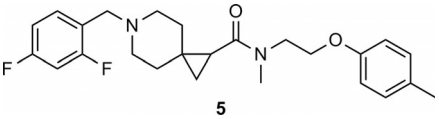
further condensed to a set of 40 purchasable compounds (Table S1 in the Supporting Information) by four methods:

- Method 1. Ten compounds were selected based on the overall highest CXCR4 prediction probability of the CNN (highest network *Score*). None of these compounds were predicted as chemokine receptor ligands by SPiDER software,[23] an independent tool for target prediction.
- Method 2. Ten compounds were selected based on the overall most confident chemokine receptor predictions made by the SPiDER target prediction software.[23]
- Method 3. The 7423 compounds with a CNN *Score* >0.995 were sorted by decreasing *Score* values and subjected to target prediction by SPiDER.[23] From this sorted list, the ten top-ranking compounds were selected that received SPiDER predictions as potential chemokine receptor ligands. Their ranks were between 17 and 95 of the sorted CNN *Score* list.
- Method 4. A scaffold tree of all 7423 screening compounds with a CNN *Score* >0.995 was generated. The subtrees containing the compounds selected by Methods 1, 2, and 3 were excluded. The remaining ten largest subtrees were prioritized and the compounds from each subtree were sorted according to their CNN *Score*. From each of the ten subtrees the highest ranking molecule was selected which possessed a framework extending the scaffold diversity of the compounds selected by Methods 1, 2, and 3.

In total, 36 of the 40 selected compounds could be purchased from commercial suppliers and biologically tested. Five of the tested compounds showed an EC_{50} below 50 μM in cell-based CXCR4-cAMP assays (Table 2, Figure S1). Compared with the endogenous agonist CXCL12 ($EC_{50}=0.31$–0.34 nM), these observed activities are weak (Figure S2). However, the effect of the most potent ligand (**1**, $EC_{50}=16$–18 μM) had a maximal agonistic effect of 128% in the cAMP assay, which is stronger than the intracellular cAMP response evoked by CXCL12 (Figure 3). Cancer cell migration was significantly stimulated by compound **1** ($p<0.0001$, Kruskal–Wallis test, Figure 4b). Superagonist **1** features a clover-like molecular structure of three aromatic rings arranged around a central positively ionizable tertiary amine. This arrangement of functional groups follows the "three-finger pharmacophore model" of protein–protein interaction (PPI) inhibitors.[28] Apparently, the deep learning model was able to implicitly identify this structural pattern in the training data. To the best of our knowledge, compound **1** is the first synthetic low-molecular-weight full CXCR4 agonist reported to date.[29]

Morpholine scaffolds, like in compounds **2** and **3**, were previously identified as CXCR4 antagonists.[30,31] Compounds **2** and **3** were also predicted as chemokine receptor ligands by SPiDER.[23] Compound **4** is the first CXCR4 modulator containing the 3,6-diazabicyclo[3.2.2]nonane-6-yl substructure which had previously been reported for orexin 1 and orexin 2 receptor ligands.[32] Compound **5** features an 6-azaspiro[2.5]octane scaffold with partial agonistic effects on CXCR4. This scaffold is not reported in ChEMBL24 as active against any of the chemokine receptors.

Table 2. Compound activities in the cAMP assay performed in CHO cells overexpressing human CXCR4. CXCR4 activity was tested for cAMP activity with CXCL12 as positive control and AMD3465 hexahydrobromide as negative control ($N=2$).

Compound	EC_{50} [μM]	Maximal effect [%]
 **1**	16–18	119–126
 **2**	37–42	86–91
 **3**	39–43[a]	89–94
 **4**	41–47[a]	44–46
 **5**	40–43	70

[a] Approximated value from limited experimental data. Dose–response curve fitting resulted in the values. Visual inspection of the curve shows approximation.

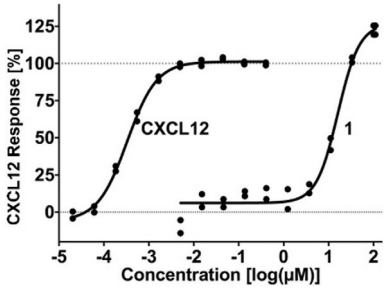


Figure 3. Concentration-dependent activation of CXCR4 by CXCL12 and compound **1** in the cAMP assay. Receptor activation is expressed as the relative effect compared with the effect of the endogenous ligand CXCL12 ($N=2$).

Regarding the compound selection method, compound **1** was on rank 11 according to the neural network *Score* (Method 1). Compounds **2**, **3**, and **5** were selected based on their high SPiDER score, ranging from 0.991 for **5** and 0.996 for **3** (Method 2). Compound **4** was among the top-ranking

© 2019 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

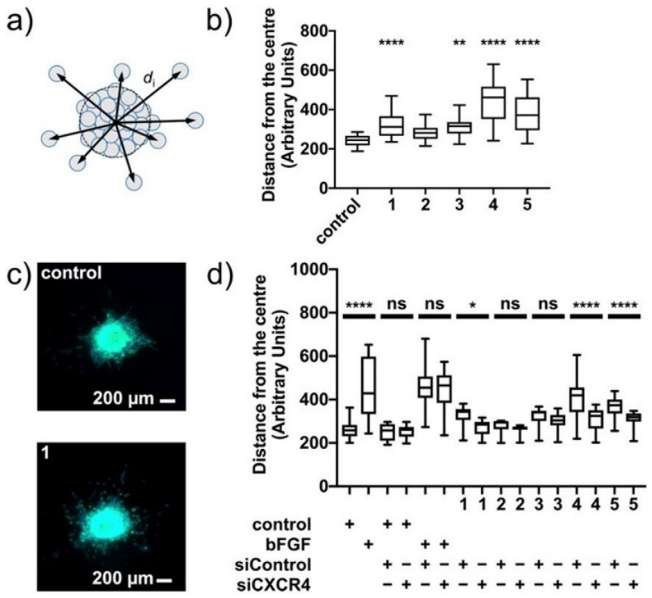


Figure 4. Spheroid invasion assay (SIA). a) Concept of the assay. Average distances d_i (arbitrary units) of cell invasion in the collagen matrix was determined by automated microscopy image analysis. b) Migration of DAOY cells after incubation in the presence of the test compounds (concentration: 10 µM) or absence of test compounds (control) after 24 h ($N=3$). The effects of compounds **1**–**5** were compared with the untreated control (** $p=0.001$–0.01, **** $p<0.0001$). Boxplots show the mean (black line), the 1st and 3rd quartiles (lines), and the 5th and 95th percentiles (whiskers). See SIA with the endogenous ligand CXCL12 in Figure S3 and evidence for CXCR4 expression in Figure S4. c) SIA fluorescence microscopy images. d) SIA with CXCR4 knockdown. Invasion of DAOY (control), DAOY with the silencing control (siControl) and DAOY with silenced CXCR4 (siCXCR4). Incubation in the absence of a stimulant (control), with bFGF (concentration: 100 ng mL^{-1}), or compounds **1**–**5** (concentration: 10 µM) for 24 h ($N=3$). The effects were compared with the corresponding silencing control (* $p=0.01$–0.05, **** $p=<0.0001$). Boxplots show the mean (black line), the 1st and 3rd quartiles (lines), and the 5th and 95th percentiles (whiskers). See evidence for CXCR4 expression in Figures S5 and S6.

molecules (rank 57) according to the neural network *Score* that were also predicted as chemokine receptor ligands by SPiDER (Method 3). No actives were retrieved based on scaffold diversity (Method 4).

For phenotypic effect screening, the 36 compounds were tested by a spheroid invasion assay (SIA, Figure 4a, b) using a cell-based model of sonic hedgehog medulloblastoma tumors, which express high levels of CXCR4 (Figure S4) and depend on CXCR4 function for tumor propagation.[33] The SIA revealed a significant invasion-activating effect of compounds **1** and **3**–**5** on DAOY medulloblastoma cells (Figure 4c). We then tested the effect of compounds **1**–**5** on cells with silenced CXCR4 (Figures 4d and S5). The silenced cells still showed invasion when stimulated with bFGF which induces cell invasion via a CXCR4 independent pathway. Comparing **1**, **4**, and **5** to their corresponding silencing control shows that invasion was significantly reduced after 24 hours of incubation. The average invasion in the CXCR4 silenced cells caused by compounds **2** and **3** was statistically insignificant.

There was no correlation between the EC_{50} values in the cAMP-based assay and cell migration in the SIA ($r^2=0.13$). To understand this discrepancy between the results of the functional and the phenotypic assays, alternative modes of action were considered. To this end, computational target predictions by SPiDER[23] and TIGER[34] software were obtained for compounds **1**–**5**. Targets with an influence on cell migration or chemotaxis, according to the respective gene ontology annotation,[35] were considered and their expression on DAOY medulloblastoma cells taken into account (Figure S5). A subset of target proteins was selected for in vitro testing (Table 3).

Table 3. Potential targets profiled for compounds **1**–**5**.[a]

Target (assay type)	Compound				
	1	**2**	**3**	**4**	**5**
CXCR4 (agonist effect)	45–53	40–64	50–57	51–58	8–38
CXCR4 (antagonist effect)	<0	<0	<0	<0	<0
CCR1 (agonist radioligand)	<0	5–8	59–64	1–2	62–66
D1 (agonist effect)	1–5	<0	<0	<0	<0
D1 (antagonist effect)	49–71	<0	<0	<0	<0
D2L (antagonist radioligand)	99–100	8–17	44–51	22–38	95–97
D2S (agonist effect)	15–22	30	80–114	30–35	45–47
D2S (antagonist effect)	44–50	n.a.	n.a.	n.a.	n.a.

[a] Values are percent (%, $N=2$) of the respective control. Compounds were tested at a single concentration of 50 µM for CCR1, D1 (D1 dopamine receptor), D2L (D2(long) dopamine receptor), and at 10 µM on CXCR4 and D2S (D2(short) dopamine receptor). CXCR4 activity was tested for cAMP activity with CXCL12 as positive control and AMD3465 hexahydrobromide as negative control. CCR1 activity was determined in a radioligand displacement assay with [^{125}I]MIP-1-α. D1 and D2S were assayed in functional assays. D1 was tested for cAMP activity, with dopamine as positive control and SCH 23390 as negative control. D2L was tested in an antagonistic radioligand assay using [^{3}H]methylspiperone. D2S was tested in an impedance assay, with dopamine as positive control and butaclamol as negative control. Compounds inducing more than 25% agonism on D2S were unsuitable for antagonism screening. n.a.: not available.

All five compounds show activity against the predicted targets (Table 3). Chemokine receptor CCR1 is a receptor found on cells of the immune system and plays a role in the signaling in inflammatory sites.[36] It is part of a larger signaling network, involving several chemokine ligands that also interact with other chemokine receptors. Dopamine D1 receptors are GPCRs that indirectly activate the adenylate cyclase, leading to elevated intracellular cAMP levels. In terms of downstream effects, D2 dopamine receptors reduce cAMP levels by inhibiting the adenylate cyclase.[37] D2 receptors exist as two alternatively spliced isoforms with different functions, the long isoform (D2L) and the short isoform (D2S). While D2L acts as a postsynaptic G_i-coupled receptor,[38] D2S receptors are presynaptic and act as D1 autoreceptors.[37] These receptors may also affect intracellular calcium levels.[39,40]

In particular, D2 dopamine receptor binding with varying preferences toward the different isoforms was observed. Compound **1** showed D1 and D2S receptor antagonistic effects and D2L binding. Whether this activity is causative for the observed effect on DAOY cell migration remains an open question at this time, because the protein expression in DAOY cells was performed on the RNA level, and no comment can be made on the isoform of the D2 receptor that is expressed in these

© 2019 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

cells. Compound **2** is a full agonist for CXCR4 according to the cAMP-based assay. However it did not invoke significant cell migration in the SIA (Figures 4d and S7). Activity testing revealed compound **2** as a weak D2S agonist, a property it shares with compounds **3**–**5**. Compounds **3** and **5** were confirmed as CCR1 and D2L ligands, with **3** being a stronger D2S agonist than **5**. These newly identified bioactivities showcase successful computational target prediction for the phenotypic screening hits.

To determine whether these results of the deep learning approach justify its computational complexity, we compared the CNN model to straightforward similarity searching with CATS, because this molecular representation also provided the input to the deep learning model. We calculated the Euclidean distance of each of the 392 known CXCR4 modulators in the training set to the screening library and sorted the library compounds by decreasing similarity to each reference ligand individually. To determine the ranks of the hits **1**–**5** according to CATS, we averaged their individual ranks in the 392 sorted lists. With this similarity approach, compound **1** was found on rank 3634337, compound **2** on rank 2860603, compound **3** on rank 3021870, and compound **4** on rank 1698982. The greatest pairwise similarity was calculated for reference CXCR4 ligand IIk[41] and hit compound **3**, which was found on average rank 53. This result suggests that none of the actives **1**–**5** would have been retrieved by CATS similarity searching alone. The CNN model enabled meaningful virtual screening and retrieved innovative bioactive compounds.

Conclusion

Drug design projects are often confronted with limited data availability, rendering the straightforward application of data-driven deep learning unfeasible. In this prospective application, we explored the usefulness of CNNs to virtual compound screening in such a situation. To enable the application of CNNs, we devised a virtual screening workflow combining the advantages of domain-specific data representation with deep model learning. This concept of combining two types of neural networks (SOM and CNN) proved successful for the given task of identifying novel CXCR4 modulators. Structurally diverse ligands were identified in a cell-based assay, and for some of the hits, a stimulating effect on cancer cell migration was confirmed. Model application led to the identification of the first reported small molecular CXCR4 agonists from a large compound collection. These bioactive compounds would not have been identified by plain similarity searching with the identical molecular representation (CATS) used for neural network modeling. Results also suggest that the observed effects of the compounds on cell migration are not or not exclusively linked to CXCR4. Preliminary activity testing points to dopamine receptors playing a role in the regulation of medulloblastoma cell migration.

Despite these encouraging results, the scope of the current approach is limited and the concept requires further development. Firstly, the model's high accuracy level in retrospective cross-validation did not translate to the prospective application. This observation is most relevant as it highlights the necessity of prospective real-world application of deep learning models in drug discovery. Theoretical retrospective estimations of a model's generalization ability might be overly optimistic. Secondly, the moderate bioactivities of the hit compounds suggest deficits of the overall modeling concept. Compared with the endogenous agonist CXCL12, the synthetic compounds are almost 50000-fold less active. However, one should consider the fact that CXL12 is a protein agonist and the hit compounds are isofunctional small molecule PPIs. This discrepancy is not a specific issue of the CNN approach but frequently observed in virtual screening. Future method developments will also have to tackle the problem of quantitative activity prediction.[42,43] However, any data-driven model can only be prospectively successful when trained with representative and reasonably accurate data sets that define chemically meaningful boundaries of the model's applicability domain. In the present example, hybrid deep network training was successful despite the comparably small number of training examples.

Experimental Section

Positive dataset: The set of CXCR4 ligands combined activity data from ChEMBL23,[17] literature sources,[44–49] and compounds from in-house projects[50] (together the "positive" set). Database entries were removed that contained 'the blockade of HIV entry" as an assay-endpoint annotation in ChEMBL, and entries without a numeric activity value or without further information on the assay conducted in the corresponding publications. Activities of different standard types (IC_{50}, EC_{50}, K_i, EC) were considered. For entries with several annotated activities, the median of the activities acquired in binding assays was considered. When several values were given, EC and values with the relation '>' were excluded and the median value calculated. Ligands were considered active if the corresponding activity value was lower or equal to 10 μM. In total, the curated positive dataset consisted of 392 molecules with reported activities on CXCR4.

Negative dataset: The negative dataset consisted of ligands from the ChEMBL23 database with non-negative activity comments, had annotated assay confidence scores >7, had 'Single Protein' as target type as well as annotations for standard metrics (IC_{50}, EC_{50}, $K_{a,b,d,e,i,m}$). Ligands annotated with the non-logarithmic activity values were discarded if they: 1) lacked a unit annotation ('standard_units' is None), 2) were tested in assays with mutated proteins, 3) had a molecular weight $<110\ g\,mol^{-1}$, and 4) were annotated as CXCR4 ligands. All valid activity data were transformed in to $\log_{10}$ units. Only ligands with log-activity values between 3 and 12 against targets other than CXCR4 were kept for model development. This data processing resulted in 495827 molecules constituting the negative dataset for CNN training.

Screening library: For virtual compound screening we compiled a pool consisting of commercially available compounds (Asinex, Delft, The Netherlands; Chembridge, San Diego, CA, USA; Enamine, Monmouth Jct., NJ, USA; Specs, Zoetermeer, The Netherlands) and an in-house virtual combinatorial library, resulting in a total of 5747961 compounds. Sanitation and standardization of all compounds was performed using MOE 2016.08[51] software. The screening compounds were encoded as descriptor arrays using CATS2 software[18] prior to CXCR4 target prediction by the hybrid model.

© 2019 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

Convolutional neural network: For the deep learning approach, a CNN was trained as binary classifier to predict ligands as either active against CXCR4 or as belonging to the negative dataset. Prior to feeding data to the CNN, a SOM with 28×28 neurons was used to cluster all bioactive molecules from ChEMBL22 according to their similarity in CATS descriptor space.[23] The activation image generated by each molecule when fed to this trained SOM was used as input data to the deep CNN implemented with Tensorflow.[52] To identify a suitable parameter set for the CNN, a grid-search over the number of hidden layers (one to three layers), types of hidden layers (convolutional layer, pooling layer, feedforward layer) and varying parameters for the different layer types was performed, resulting in 108 network topologies for evaluation. Convolutional layers had a receptive field size of 4×4 pixels with a stride of 1, padded with periodic boundaries. The only variable parameter of the convolutional layers was the number of filters used (the convolutional layer's "depth"), which could be either 40, 50, or 60. A restriction was placed on network architectures containing several convolutional layers, in that the depth from one convolutional layer to another had to increase. Pooling layers always used max-pooling with a pooling window size of 2×2 pixels and a stride of 1. Feedforward layers could contain either 50, 100 or 200 neurons. All layers were trained with rectifier linear unit (ReLU) activation, apart from the output layer which used *softmax* activation.

Each architecture was trained for 1000 epochs using fivefold cross-validation with a 4:1 ratio of training to test data. The different architectures were trained using stochastic gradient descent with a batch size of 256 and a stepwise decaying learning rate, starting at 0.007 with a decay factor of 0.9 every 40 epochs. Because the number of negative molecules exceeded our positive dataset by several orders of magnitude, negative subsampling was used for each fold by using the positive dataset and adding an equally large, randomly selected amount of ligands from the negative dataset. Cross-entropy with an L2 regularization factor of 0.01 was chosen as error function for CNN training, performed with a momentum (*momentum* = 0.9) optimizer. Among the 108 architectures tested, the architecture displaying the highest average Matthews correlation coefficient (MCC)[53] during cross-validation was trained on all the data and then used for virtual screening of CXCR4 active ligands in the screening library.

The network models were implemented in Python (2.7) using the libraries TensorFlow (1.4.0),[52] scikit-learn (0.19.1),[54] scipy (1.0.0),[55] numpy (1.13.3),[56] matplotlib (2.1.1),[57] and pandas (0.21.0).[58] The deep-learning calculations were performed on a Linux-computer (Ubuntu 16.04) with four 3.6 GHz Intel Core i7 6850 K processors and an ASUS GeForce GTX1080Ti STRIX graphics card. Further calculations and analysis were performed on Macintosh Workstation (OS X Yosemite, 10.10.5, 2×2.26 GHz Quad-Core Intel Xeon with 48GB memory). MOE 2016.08[50] was used for the standardization of molecules. Analyses were performed using Python (3.5.3) with seaborn (0.7.1), matplotlib (2.0.0)[57] Pandas (0.19.3),[58] RDKit (2016.03.4),[59] and libraries. SPiDER predictions[23] were calculated using Knime (2.12.02).

In silico target prediction with SPiDER: The SPiDER software predicts potential targets for given molecule queries.[23] The input molecular structures were standardized for pH 7 using MOE2016.08.[51] CATS descriptors[18] and MOE descriptors were calculated and used as input for SPiDER.[23] Only the target predictions with p values < 0.05 were considered.

Scaffold diversity: The diversity of generated scaffolds was analyzed using Scaffold Hunter software (scaffold-hunter-2.6.3).[60]

In vitro tests: Compounds selected for testing were ordered from the respective suppliers and stock concentrations of 10 mM in DMSO prepared. The stock solutions were used for subsequent testing.

Functional assay for CXCR4: Functional assessment of the compound was determined by antagonistic and agonistic effect at 10 μM in a cAMP assay. CXCR4 was tested for cAMP activity with CXCL12 as positive control and AMD 3465 hexahydrobromide as negative control. Where the efficacy was higher than 50%, dose–response studies were conducted. The assays were performed by DiscoverX (Fremont, CA, USA) on a fee-for-service basis.

Silencing CXCR4 using RNA interference: DAOY cells with a confluence of approximately 75% of total surface area were transfected with small interference RNA (siRNA) specific for CXCR4 (assay ID: s15412, Thermofischer Scientifc) or Silencer select negative control (assay ID: 4390843, Ambion). The siRNAs were used at the final concentration of 5 nmol. The transfection was facilitated using DharmaFECT 4 transfection reagent (T-2004-03, Dharmacon). After 48 hours, RNA and proteins were isolated from DAOY cells. The gene expression and protein expression were determined by qPCR and Immunoblotting, respectively. Upon successful downregulation of CXCR4, the transfected cells were used for SIA.

Spheroid invasion assay (SIA): The effect of compounds on cell invasion and the phenotype of the medulloblastoma tumor cell line were determined. DAOY cells were tested by SIA, as described.[61] DAOY cells stably expressing lifeact (LA) enhanced green fluorescent protein (EGFP) produced by lentiviral transduction with pLenti-LA-EGFP were used for SIA. In brief, 1000 DAOY LA-EGPF cells per 100 μL per well were seeded in a 96-well cell-repellent microplate (650790, Greiner Bio-One). The cells were incubated overnight at 37 °C to form spheroids. 70 μL of the medium was removed from each well, and remaining medium with spheroid was overlaid with 2.5% (final concentration) of ice-cold bovine collagen 1 (5005-B, Advanced BioMatrix, San Diego, CA, USA). The collagen was allowed to polymerize for 1 h at 37 °C. Following the polymerization of collagen, fresh serum free medium was added to the cells and then treated with 10 μM (final concentration) of the compounds. The embedded cells were allowed to invade the 3D collagen matrix for 24 h, after which they were fixed with 4% PFA and stained with Hoechst. Images were acquired on an Axio Observer 2 mot plus fluorescence microscope (Zeiss, Munich, Germany) using a 5× objective. The extent of cell invasion was determined as the average of the distance invaded by the cells from the center of the spheroid, using automated cell dissemination counter.[52]

Screening assays for other targets: Functional assessment of the compounds was determined at 10 μM. Radioligand displacement assay with [^{125}I]MIP-1-α was performed for CCR1. D1 and D2S were assayed in functional assays. D1 was tested for cAMP activity, with dopamine as positive control and SCH 23390 as negative control. D2L was tested in an antagonistic radioligand assay using [^{3}H]methylspiperone. D2S was tested in an impedance assay, with dopamine as positive control and butaclamol as negative control. Compounds inducing more than 25% agonism on D2S were unsuitable for antagonism screening. The assays were performed by Cerep (Celle l'Evescault, France) on a fee-for-service basis.

Statistical analysis. All statistical analyses were as performed using Prism v7 on Macintosh (GraphPad Software, La Jolla, CA, USA). Results of the phenotypic assay were tested on their intrinsic variability by one-way ANOVA (Kruskal–Wallis test). Compounds that show no significant intrinsic variability were compared with the control using Dunn's multiple comparisons test.

© 2019 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

Acknowledgements

This research was supported financially by the Swiss National Science Foundation (SNF grant number CR32I2_159737 to G.S.).

Conflict of Interest

The authors declare the following competing interests: G.S. declares a potential financial conflict of interest in his role as life science industry consultant and cofounder of inSili.com GmbH, Zürich.

Keywords: artificial intelligence · convolutional neural network · drug discovery · medulloblastoma · virtual screening

[1] A. Krizhevsky, I. Sutskever, G. E. Hinton, *Commun. ACM* **2017**, *60*, 84 – 90.
[2] E. Gawehn, J. A. Hiss, G. Schneider, *Mol. Inf.* **2016**, *35*, 3 – 14.
[3] O. Z. Kraus, J. L. Ba, B. J. Frey, *Bioinformatics* **2016**, *32*, i52 – i59.
[4] J. T. Cuperus, B. Groves, A. Kuchina, A. B. Rosenberg, N. Jojic, S. Fields, G. Seelig, *Genome Res.* **2017**, *27*, 2015 – 2024.
[5] T. Kohonen, *Neural Networks* **2013**, *37*, 52 – 65.
[6] K. Goldberg, *Nat. Mach. Intell.* **2019**, *1*, 2 – 4.
[7] P. M. Murphy, *Annu. Rev. Immunol.* **1994**, *12*, 593 – 633.
[8] J. B. Rubin, *Semin. Cancer Biol.* **2009**, *19*, 116 – 122.
[9] D. Rodríguez, H. Gutiérrez-de-Terán, *Proteins Struct. Funct. Bioinf.* **2012**, *80*, 1919 – 1928.
[10] K. E. Luker, M. Gupta, G. D. Luker, *FASEB J.* **2009**, *23*, 823 – 834.
[11] Q. Ma, D. Jones, P. R. Borghesani, R. A. Segal, T. Nagasawa, T. Kishimoto, R. T. Bronson, T. A. Springer, *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 9448 – 9453.
[12] Y. Feng, C. C. Broder, P. E. Kennedy, E. A. Berger, *Science* **1996**, *272*, 872 – 877.
[13] J. A. Burger, T. J. Kipps, *Blood* **2006**, *107*, 1761 – 1767.
[14] V. M. Truax, H. Zhao, B. M. Katzman, A. R. Prosser, A. A. Alcaraz, M. T. Saindane, R. B. Howard, D. Culver, R. F. Arrendale, P. R. Gruddanti, T. J. Evers, M. G. Natchus, J. P. Snyder, D. C. Liotta, L. J. Wilson, *ACS Med. Chem. Lett.* **2013**, *4*, 1025 – 1030.
[15] Z. Li, Y. Wang, C. Fu, C. X. Wang, J. J. Wang, Y. Zhang, D. Zhou, Y. Zhao, L. Luo, H. Ma, W. Lu, J. Zheng, X. Zhang, *Eur. J. Med. Chem.* **2018**, *149*, 30 – 44.
[16] Y. Yoshikawa, K. Kobayashi, S. Oishi, N. Fujii, T. Furuya, *Bioorg. Med. Chem. Lett.* **2012**, *22*, 2146 – 2150.
[17] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, *Nucleic Acids Res.* **2012**, *40*, D1100 – D1107.
[18] M. Reutlinger, C. P. Koch, D. Reker, N. Todoroff, P. Schneider, T. Rodrigues, G. Schneider, *Mol. Inf.* **2013**, *32*, 133 – 138.
[19] T. Kohonen, *Biol. Cybern.* **1982**, *43*, 59 – 69.
[20] S. Anzali, J. Gasteiger, U. Holzgrabe, J. Polanski, J. Sadowski, A. Teckentrup, M. Wagener in *3D QSAR in Drug Design, Vol. 2* (Eds.: H. Kubinyi, et al.), Kluwer, **1998**, pp. 273 – 299.
[21] P. Schneider, A. T. Müller, G. Gabernet, A. L. Button, G. Posselt, S. Wessler, J. A. Hiss, G. Schneider, *Mol. Inf.* **2017**, *36*, 1600011.
[22] A. Afantitis, G. Melagraki, P. A. Koutentis, H. Sarimveis, G. Kollias, *Eur. J. Med. Chem.* **2011**, *46*, 497 – 508.
[23] D. Reker, T. Rodrigues, P. Schneider, G. Schneider, *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 4067 – 4072.
[24] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors, Vols. 1 and 2* (Eds.: R. Todeschini, V. Consonni) Wiley-VCH, Weinheim, **2008**.
[25] B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum, *Science* **2015**, *350*, 1332 – 1338.
[26] H. Altae-Tran, B. Ramsundar, A. S. Pappu, V. Pande, *ACS Cent. Sci.* **2017**, *3*, 283 – 293.
[27] G. Schneider, P. Schneider, *Expert Opin. Drug Discovery* **2017**, *12*, 271 – 277.
[28] A. Dömling, *Curr. Opin. Chem. Biol.* **2008**, *12*, 281 – 291.
[29] M. Lefrançois, M. R. Lefebvre, G. Saint-Onge, P. E. Boulais, S. Lamothe, R. Leduc, P. Lavigne, N. Heveker, E. Escher, et al., *ACS Med. Chem. Lett.* **2011**, *2*, 597 – 602.
[30] A. Zhu, W. Zhan, Z. Liang, Y. Yoon, H. Yang, H. E. Grossniklaus, J. Xu, M. Rojas, M. Lockwood, J. P. Snyder, D. C. Liotta, H. Shim, et al., *J. Med. Chem.* **2010**, *53*, 8556 – 8568.
[31] K. S. Gudmundsson, P. R. Sebahar, L. D. Richardson, J. F. Miller, E. M. Turner, J. G. Catalano, A. Spaltenstein, W. Lawrence, M. Thomson, S. Jenkinson, et al., *Bioorg. Med. Chem. Lett.* **2009**, *19*, 6399 – 6403.
[32] P. J. Coleman, J. D. Schreier, A. J. Roecker, S. P. Mercer, G. B. McGaughey, C. D. Cox, G. D. Hartman, C. M. Harrell, D. R. Reiss, S. M. Doran, S. L. Garson, W. B. Anderson, C. Tang, T. Prueksaritanont, C. J. Winrow, J. J. Renger, et al., *Bioorg. Med. Chem. Lett.* **2010**, *20*, 4201 – 4205.
[33] S. A. Ward, N. M. Warrington, S. Taylor, N. Kfoury, J. Luo, J. B. Rubin, et al., *Cancer Res.* **2017**, *77*, 1416 – 1426.
[34] P. Schneider, G. Schneider, *Angew. Chem. Int. Ed.* **2017**, *56*, 11520 – 11524; *Angew. Chem.* **2017**, *129*, 11678 – 11682.
[35] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, The Gene Ontology Consortium, *Nat. Genet.* **2000**, *25*, 25 – 29.
[36] P. P. Tak, A. Balanescu, V. Tseluyko, S. Bojin, E. Drescher, D. Dairaghi, S. Miao, V. Marchesin, J. Jaen, T. J. Schall, P. Bekker, *Ann. Rheum. Dis.* **2013**, *72*, 337 – 344.
[37] J.-M. Beaulieu, R. R. Gainetdinov, *Pharmacol. Rev.* **2011**, *63*, 182 – 217.
[38] A. Usiello, J. H. Baik, F. Rougé-Pont, R. Picetti, A. Dierich, M. LeMeur, P. V. Piazza, E. Borrelli, *Nature* **2000**, *408*, 199 – 203.
[39] A. J. Rashid, C. H. So, M. M. C. Kong, T. Furtak, M. El-Ghundi, R. Cheng, B. F. O'Dowd, S. R. George, *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 654 – 659.
[40] A. L. Frederick, H. Yano, P. Trifilieff, H. D. Vishwasrao, D. Biezonski, J. Mészáros, E. Urizar, D. R. Sibley, C. Kellendonk, K. C. Sonntag, D. L. Graham, R. J. Colbran, G. D. Stanwood, J. A. Javitch, *Mol. Psychiatry* **2015**, *20*, 1373 – 1385.
[41] R. Bai, Z. Liang, Y. Yoon, S. Liu, T. Gaines, Y. Oum, Q. Shi, S. R. Mooring, H. Shim, *Eur. J. Med. Chem.* **2016**, *118*, 340 – 350.
[42] G. Schneider, *Nat. Rev. Drug Discovery* **2010**, *9*, 273 – 276.
[43] G. Schneider, *Nat. Mach. Intell.* **2019**, *1*, 128 – 130.
[44] M. M. Mysinger, D. R. Weiss, J. J. Ziarek, S. Gravel, A. K. Doak, J. Karpiak, N. Heveker, B. K. Shoichet, B. F. Volkman, *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 5517 – 5522.
[45] D. Das, K. Maeda, Y. Hayashi, N. Gavande, D. V. Desai, S. B. Chang, A. K. Ghosh, H. Mitsuya, *Antimicrob. Agents Chemother.* **2015**, *59*, 1895 – 1904.
[46] S. Oishi, T. Kuroyanagi, T. Kubo, N. Montpas, Y. Yoshikawa, R. Misu, Y. Kobayashi, H. Ohno, N. Heveker, T. Furuya, N. Fujii, *J. Med. Chem.* **2015**, *58*, 5218 – 5225.
[47] Z. G. Zachariassen, S. Thiele, E. A. Berg, P. Rasmussen, T. Fossen, M. M. Rosenkilde, J. Våbenø, B. E. Haug, *Bioorg. Med. Chem.* **2014**, *22*, 4759 – 4769.
[48] Z. G. Zachariassen, S. Karlshøj, B. E. Haug, M. M. Rosenkilde, J. Våbenø, *J. Med. Chem.* **2015**, *58*, 8141 – 8153.
[49] H. Ha, B. Debnath, S. Odde, T. Bensman, H. Ho, P. M. Beringer, N. Neamati, *J. Chem. Inf. Model.* **2015**, *55*, 1720 – 1738.
[50] D. Reker, P. Schneider, G. Schneider, *Chem. Sci.* **2016**, *7*, 3919 – 3927.
[51] *Molecular Operating Environment (MOE), 2013.08*, Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, H3A 2R7 (Canada), **2018**.
[52] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, Preprint at: https://arxiv.org/pdf/1603.04467.pdf, **2015**.
[53] B. W. Matthews, *Biochim. Biophys. Acta Protein Struct.* **1975**, *405*, 442 – 451.
[54] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825 – 2830.

© 2019 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

[55] E. Jones, T. Oliphant, P. Peterson, et al., *SciPy: Open Source Scientific Tools for Python*, **2001**: http://www.scipy.org (last accessed October 9, 2019).
[56] S. van der Walt, S. C. Colbert, G. Varoquaux, *Comput. Sci. Eng.* **2011**, *13*, 22 – 30.
[57] J. D. Hunter, *Comput. Sci. Eng.* **2007**, *9*, 90 – 95.
[58] W. McKinney, *Data Structures for Statistical Computing in Python in Proceedings of the 9th Python in Science Conference* (Eds.: S. van der Walt, J. Millman), **2010**, pp. 51 – 56.
[59] G. Landrum, *RDKit: Open-Source Cheminformatics Software*, **2016**: http://www.rdkit.org (last accessed October 9, 2019).
[60] S. Wetzel, K. Klein, S. Renner, D. Rauh, T. I. Oprea, P. Mutzel, H. Waldmann, *Nat. Chem. Biol.* **2009**, *5*, 581 – 583.
[61] K. S. Kumar, M. Pillong, J. Kunze, I. Burghardt, M. Weller, M. A. Grotzer, G. Schneider, M. Baumgartner, *Sci. Rep.* **2015**, *5*, 15338.

Manuscript received: May 27, 2019
Revised manuscript received: August 7, 2019
Accepted manuscript online: August 16, 2019
Version of record online: November 14, 2019

© 2019 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim